针对 RDF 概率图查询的基数估计方法

章登义,吴文李,欧阳黜霏
(武汉大学计算机学院,湖北武汉 430072)

摘 要: 资源描述框架图查询中,准确估计查询结果的大小是查询优化器中的关键步骤.已有方法忽略了该图自身的不确定性以及子查询间的关联关系,无法有效估计结果.针对该问题,本文提出一种基于贝叶斯模型的基数估计方法.该方法引入贝叶斯网络模型,挖掘出子查询内的属性依赖.同时,在这些属性依赖的基础上提出子网拼接方法,计算出子查询间的影响因子.最后,利用以上信息准确估计出任意查询结果集的基数.实验表明:与已有方法相比,本文方法的准确性提高15%以上,性能没有大幅度下降.

关键词: 不确定资源描述框架图; 查询处理; 选择基数估计; 查询优化
中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112 (2015)09-1745-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.09.010

Cardinality Estimation in Query for Probability RDF Graphs

ZHANG Deng-yi, WU Wen-li, OUYANG Chu-fei
(*School of Computer, Wuhan University, Wuhan, Hubei* 430072, *China*)

Abstract: In RDF(Resource Description Framework) graph query, accurately estimating the size of the query result is a crucial step to the query optimizer. The previous work, which ignores both the uncertainty of RDF graph itself and the correlations between subqueries, is difficult to obtain accurate estimations. To solve this problem, this paper proposes an estimation method based on Bayesian probability model. Our method introduces Bayesian network model for subqueries to dig out the dependencies between properties in subqueries. At the meanwhile, based on these dependencies we propose a connection approach of subnets to compute the impact factors between subqueries. Finally, we exploit the above information to accurately estimate the cardinality of the result about an arbitrary query. The experiments indicate that the accuracy of our estimation results is improved by over 15% and that the query run-time is not increased significantly in comparison with the previous art.

Key words: uncertain RDF graph; query processing; selectivity estimation; query optimization

1 引言

资源描述框架的三元组数据形式为<资源,属性,描述>(〈subject, property, object〉).针对该图数据查询做选择基数估计的方法[1~4]约分为三类:直方图法[1]、字符集法[2]和概率估计法[3,4].后两者的准确性要比直方图法的高,但建立在各个子查询相互独立的假设之上,与实际相互关联的查询图相悖,因此估计结果偏差过大.

同时,利用文献[5]的概率模型对确定数据进行建模,建模后的输出包含大量的不确定顶点,同一资源的一条边指向多个不同的值或者多个不同的其他资源,即上述方法均忽略了RDF数据的不确定性[5,6].因此,已有方法无法得到准确的估计值.

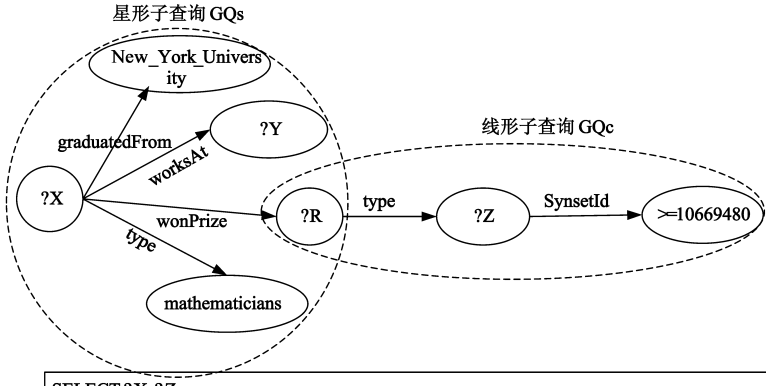
针对于此,本文既考虑RDF图数据本身的不确定性同时又不忽略子查询间的相互关联,准确估计出任意查询结果的大小.本文工作如下:(1)为概率图子查询引入贝叶斯网络[7]以挖掘其中的属性依赖,形成贝叶斯子网;(2)提出子网拼接方法以计算子查询间的影响因子,完成对任意查询的选择基数估计.

2 问题定义

定义1 RDF概率图用三元组($V(G)$, $E(G)$, $S(G)$)表示,其中,有:$V(G)$是顶点v_i的有限集,每个元素都附有概率标签$l(v_i)$;$E(G)$是边e_{ij}的有限集,每个元素都附有一个标签$l(e_{ij})$;$S(G)$是一个有限集,包含概率标签的具体信息.

定义2 给定RDF概率图G',查询语句Q,以及临界值$a\in[0,1)$,概率图查询的定义为:找出满足条件(1)和(2)的任意匹配子图$S_{Qi}\in(S_{Q1}, S_{Q2}, \cdots, S_{Qn})$,$0<i<n+1$.其中,$(S_{Q1}, S_{Q2}, \cdots, S_{Qn})\in G'$,匹配条件为:(1)

收稿日期:2014-05-23;修回日期:2014-10-08;责任编辑:李勇锋




```
SELECT ?X ?Z
 WHERE {
   ?X  graduatedFrom  'New _York_University'.   ?X  worksAt      ?Y.
   ?X  type           'mathematicians'.         ?X  wonPrize     ?R.
   ?R  type           ?Z.                       ?Z  hasSynsetId  '>=10669480' }
```

图1 针对YAGO且a=0.34的SPARQL查询

S_{Qi}与 Q 图模式匹配;(2) S_{Qi}的出现概率 $P(S_{Qi}) > a$.

定义 2 查询语句上表现为带 a 的 SPARQL(Simple Protocol and RDF Query Language)查询,如图 1 所示.

定义 3 概率图子查询表示为 $Q_i \in (Q_1, Q_2, \cdots, Q_n), 0 < i < n+1$.其中,子查询集$(Q_1, Q_2, \cdots, Q_n)$是由带 a 的 SPARQL 查询图 Q 划分后所得的集合.

子查询 Q_i有两种类型,分别是星形子查询 Q_s和线形子查询 Q_c.子查询用属性描述,图 1 的子查询分别为 Q_s'(graduatedFrom, worksAt, wonPrize, type)和 Q_c'(type, synsetId).

3 查询的基数估计

3.1 贝叶斯网络和条件概率表

贝叶斯网络[8]由两个部分组成,分别为有向无环图和条件概率表.有向无环图的每个节点代表一个随机变量,每条边表示一个概率依赖.每个变量对应一个条件概率表 CPT(Conditional Probability Table).每个 CPT 对应一个条件分布.假设 $X = (x_1, x_2, \cdots, x_n)$是变量性 $Y_1, Y_2, \cdots, Y_n$的元组数据.贝叶斯网络下的联合概率为:

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} (P(x_i \mid \text{Parents}(Y_i))) \quad (1)$$

其中,$P(x_1, x_2, \cdots, x_n)$表示 X 取某一组合值的概率,而 $P(x_i \mid \text{Parents}(Y_i))$对应于 CPT$(Y_i)$的表目.

3.2 RDF 概率图查询的贝叶斯网和条件概率表

3.2.1 基础表的构造

构造基础表时,通过观察分析 RDF 图数据发现,所有的三元组以组为单位对资源进行描述,即多个三元组同时描述一个资源.因此,对于数据集 R 中的每个资源 s,都有紧密相关的字符集:

$$\text{Sc}(s) = \{p \mid \exists\, o : (s, p, o \in R)\} \quad (2)$$

式(2)涵盖了 s 所有的 1-hop 星状关联属性和资源.利用该公式为图 1 的星形子查询获取基础表,结果如表 1 所示.表 1 的每条记录表示一个人的信息,每个人都包含 Sc(? X)中的 4 个属性.

表 1 Sc(? X) = $\{p \mid \exists\, o : (?\ X, p, o \in R)\}$基础表

? X	graduatedFrom	worksAt	type	wonPrize
X_1	'Punjab University'	'ScientificInstitut' × 0.97	'ShanghaiPeople'	'Hilal Imtiaz'
X_1	'Punjab University'	'EnergyCommission' × 0.03	'ShanghaiPeople'	'Hilal Imtiaz'
X_2	'New York University'	'EnergyCommission'	'Muslim scholars'	'Fields Medal' × 0.3
…	…	…	…	…

同理,RDF 图数据中,所有的三元组亦可按属性进行分组并描述不同资源,即属性(property)相同的一组三元组集.因此,给定属性 p,有集合:

$$\text{Sc}(p) = \{s, p, o \mid \exists\ S, \exists\ o : (s, p, o) \in R\} \quad (3)$$

或 $\neg$ $\text{Sc}(p) = \{o, p, s \mid \exists\ S, \exists\ o : (s, p, o) \in R\}$ (4)

其中,式(3)的顺序是资源、属性、描述,即 SPO,而式(4)的顺序是 OPS.因此,Q_c'的基础表为 $\text{Sc}(p_2) \cup \neg\ \text{Sc}(p_1)$.

本文通过对数据进行预处理以实现基础表的快速构建.因此,复杂度有两部分:一是预处理的复杂度;二是线上开销.预处理部分,计算复杂度均为 $O(n)$.而线上构造时,星形子查询的复杂度为 $\log(n)$.线形子查询的开销为 $O(m \cdot n \cdot \log(n))$,即 $O(n \cdot \log(n))$.

3.2.2 子查询的贝叶斯网构造

由文献[4]得到启发,本文利用文献[9]的条件独立

CI 测试算法(Conditional Independence Test)判断属性与属性间的依赖关系.其中,引入相互信息以计算两属性的关联.给定变量 X 和 Y,X 和 Y 的相互信息 $I(X,Y)$ 的定义为:

$$I(X,Y)=\sum_{x,y}\Pr(x,y)\log\frac{\Pr(x,y)}{\Pr(x)\cdot\Pr(y)} \tag{5}$$

条件相互信息的定义为:

$$I(X,Y\mid M)=\sum_{x,y,m}\Pr(x,y,m)\log\frac{\Pr(x,y\mid m)}{\Pr(x\mid m)\cdot\Pr(y\mid m)} \tag{6}$$

其中,M 表示某一基础表的属性集.当 $I(X,Y\mid M)$ 小于临界值 ε,X 和 Y 在 M 条件下有向分离(d-separated).

贝叶斯网络构造算法包括三个步骤.步骤一主要计算属性对之间的相互信息以衡量两者关联程度,利用该信息画出初步的依赖图.图中不包含任何环.步骤二主要在当前的图上为非有向分离的属性对添加依赖边.而步骤三利用 CI 测试检验图中的所有边,移除有向分离边并确定依赖边的方向.构造算法中的参数 ε,取 $\varepsilon<0.05$.

对基础表进行贝叶斯学习,子查询 Q_s'(graduatedFrom, worksAt, wonPrize, type)和 Q_c'(type, synsetId)的贝叶斯网分别如图 2(1)和图 2(2)所示.

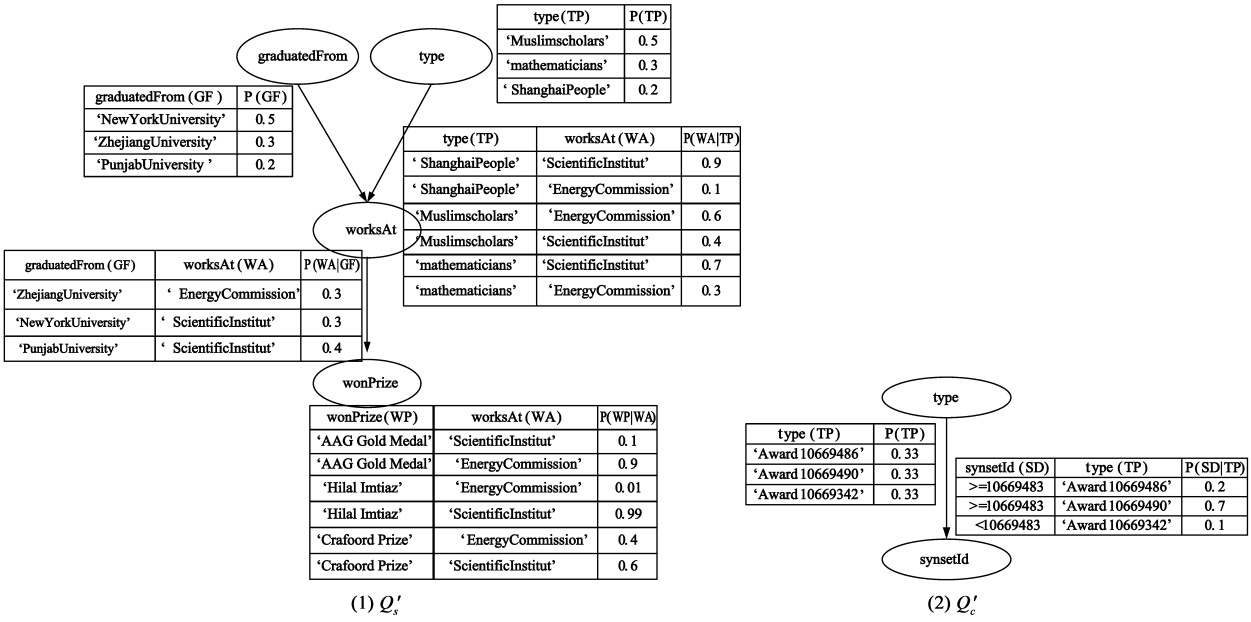
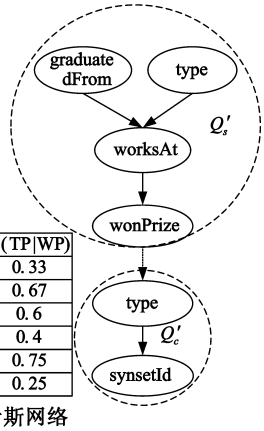


图 2　针对图1所构造的子查询贝叶斯网及其CPT

3.2.3　子网拼接

本节为子查询的相邻属性构造基础表,并在此基础上进行贝叶斯网络学习以确定依赖边的方向,然后将此依赖边添加至子查询间,以完成子网的拼接.拼接后,本节得到整个查询的属性依赖关系图.



wonPrize(WP)	type(TP)	P(TP|WP)
'Hilal Imtiaz'	'Award10669486'	0.33
'Hilal Imtiaz'	'Award10669490'	0.67
'Fields Medal'	'Award10669342'	0.6
'Fields Medal'	'Award10669372'	0.4
'Crafoord Prize'	'Award10669347'	0.75
'Crafoord Prize'	'Award10669309'	0.25

图3　$Q_s'\cup Q_c'$的贝叶斯网络

图 2 进行子网拼接时,本文利用内存中已有表格构造基础表 $\mathrm{Sc}(Q_c'=\text{'type'})\cup\neg\,\mathrm{Sc}(Q_s'=\text{'wonPrize'})$,并进行贝叶斯学习,学习结果如图 3 所示.图 3 中,子网拼接步骤为 Q_s' 和 Q_c' 之间添加了依赖 wonPrize→type,并对 type 的 CPT 进行相应的更新.因为 type 不再独立于 wonPrize.更新后,type 的 CPT 如图 3 所示.

3.2.4　基数估计的计算过程

本节的计算过程包含两部分,第一部分在不考虑数据不确定性的情况下,估算最终结果的大小.第二部分主要根据数据的不确定性对第一部分的估计值进行修正.

第一部分计算子查询间的影响因子并利用该因子估计查询结果的基数.图 1 的 SPARQL 查询 Q',估计结果 $\mathrm{card}(Q')$ 的推导过程如下:

$$\begin{aligned}\mathrm{card}(Q') &= \mathrm{sel}(Q')\cdot\mid R\mid\\ &=\mid R\mid\cdot P(Q_s')\,u\cdot P(Q_c')\end{aligned}$$

$$= |R| \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} P(Q'_{s_i}) \cdot P(Q'_{c_j} \mid Q'_{s_j}) \tag{7}$$

$$\begin{aligned} P(Q_s) &= P(\text{graduatedFrom}, \text{type}, \text{worksAt}, \text{wonPrize}) \\ &= P(\text{graduatedFrom}) \cdot P(\text{type}) \cdot P(\text{worksAt} \mid \text{type}) \\ &\quad \cdot P(\text{worksAt} \mid \text{graduatedFrom}) \\ &\quad \cdot P(\text{wonPrize} \mid \text{worksAt}) \end{aligned} \tag{8}$$

$$\begin{aligned} P(Q'_c \mid Q'_s) &= P(\text{type}, \text{synsetId} \mid \text{wonPrize}) \\ &\quad \cdot P((\text{synsetId} \mid \text{type}) \mid \text{wonPrize}) \\ &= P(\text{type}, \text{synsetId} \mid \text{wonPrize}) \\ &\quad \cdot P(\text{synsetId} \mid \text{wonPrize}) \end{aligned} \tag{9}$$

其中，$|R|$为表 1 和表 2 聚合后的总记录数，i 和 j 的值表示因变量？Z、？R 或？Y 的一组属性实例.

第二部分，本文在计算 $\text{sel}(Q') = \sum_{i=1}^{n} \sum_{j=1}^{m} P(Q'_{s_i}) \cdot P(Q'_{c_j} \mid Q'_{s_i})$ 的过程中，获取每个匹配子图的出现概率，一旦出现 $P(Q'_{s_i}) \cdot P(Q'_{c_j} \mid Q'_{s_i}) < a$，默认将该实例图的出现概率置零.

4 实验与评价

4.1 实验设置

本实验所用数据集分别为：LUBM，DBLP 和 YAGO. 同时，本节分别为三个数据集设计查询. 实验对照对象有基于字符特征的估计方法[2]，记为 CS 以及基于概率框架的选择基数估计[4]，记为 PF. 本文方法记为 NCN.

4.2 准确性评价

本节实验主要测试关联子查询对估计准确性的影响. 实验结果如图 4. 图中，X 坐标轴表示查询语句集，Y 轴表示相对误差 RE(relative error)的值. 其中，RE(sel, sel′) = |sel-sel′|/sel，sel(Q)是真实值，sel′(Q)是估计值.

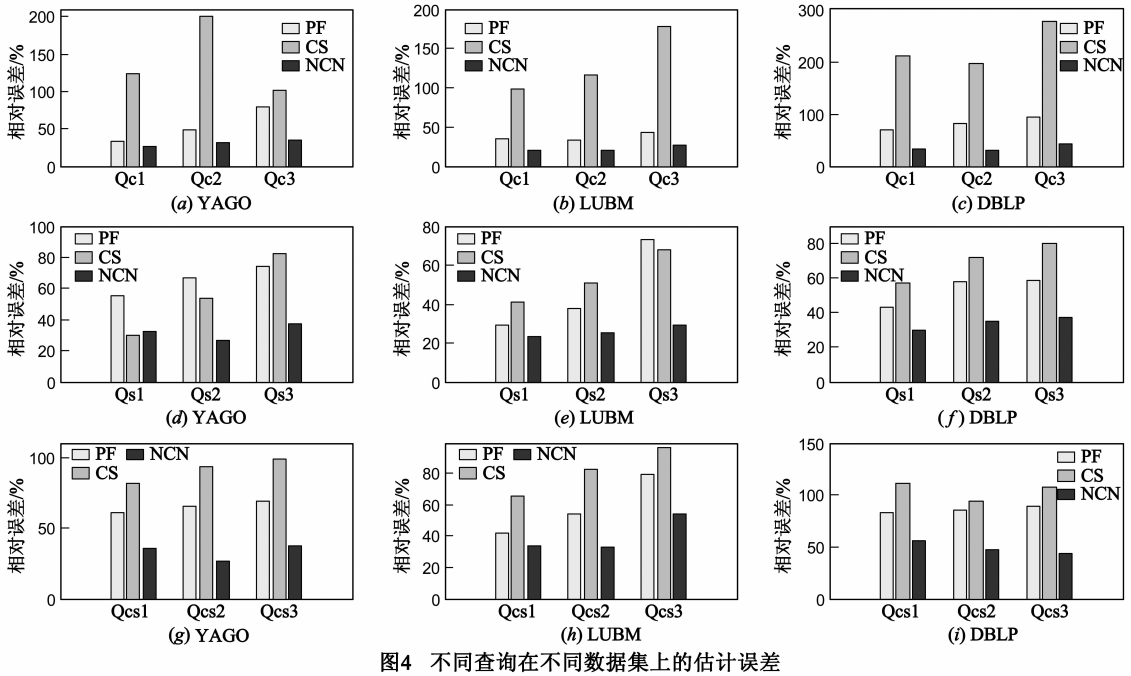


图4 不同查询在不同数据集上的估计误差

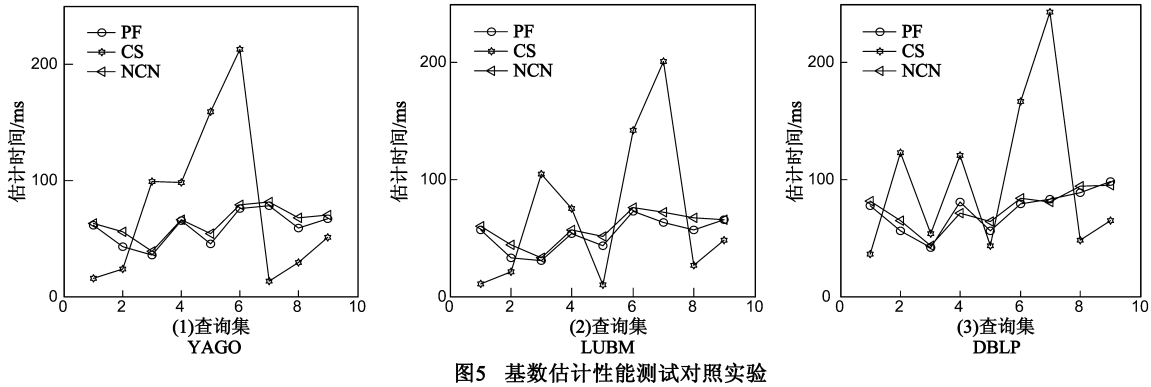


图5 基数估计性能测试对照实验

图 4(*a*)、4(*b*)、4(*c*)是线形查询上的对照实验.由于 NCN 和 PF 均用严谨的数学方法计算估计值,而 CS 仅通过对部分相连属性进行预存储以获取结果,因此,NCN 和 PF 的准确性均高于 CS 的.同时,NCN 不仅考虑子查询内的属性依赖,还准确计算子查询间的影响,所以,NCN 的准确性最高.图 4(*d*)、4(*e*)、4(*f*)的星形查询中,NCN 在准确性上仍高于其他两种办法.但是,CS 的准确性与 PF 的几乎持平,甚至高于后者.该现象的主要原因在于 CS 主要针对星形查询而设计,因此,CS 的准确性比 PF 的高.最后的一组图的结果几乎是前面综合.

4.3 性能评价

本组实验分别测试 3 种查询在不同数据集上分别使用 PF、CS 和 NCN 所需的估计时间,结果如图 5 所示.

图 5 中,CS 所需的时间在一定范围内随机出现,时高时低,NCN 与 PF 测试结果几乎同一曲线且 NCN 时间略长.与 NCN 高出 PF 的准确性相比,NCN 的估计时间是可接受的.

5 结论

本文为不确定概率图查询提供了可用于查询优化的基数估计方法.对于任意查询,我们引入贝叶斯模型,提出模型拼接方法以计算子查询间的影响因子,同时,利用该系数完成对查询的选择基数估计.最后,我们设计实验以评价本文方法.实验表明,本文方法高效准确完成不确定图查询的选择基数估计.

参考文献

[1] Yannis E I. The history of histograms (abridged)[A]. International Conference on Very Large Data Bases[C]. San Fransisco: Morgan Kaufmann, 2003. 19 – 30.

[2] Thomas N, Guido M. Characteristic sets: accurate cardinality estimation for RDF queries with multiple joins[A]. IEEE International Conference on Data Engineering[C]. Washington: IEEE Press, 2011. 984 – 994.

[3] Markus S, et al. SPARQL basic graph pattern optimization using selectivity estimation[A]. International World Wide Web Conferences[C]. London: Springer, 2008. 595 – 604.

[4] Hai H, Chengfei L. Estimating selectivity for joined RDF triple patterns[A]. ACM International Conference on Information and Knowledge Management[C]. New York: ACM, 2011. 1435 – 1444.

[5] Xiang L, Lei C. Efficient query answering in probabilistic RDF graphs[A]. ACM Conference on Management of Data[C]. New York: ACM, 2011. 157 – 168.

[6] Ye Y, et al. Efficient subgraph similarity search on large probabilistic graph databases[J]. The Proceedings of the VLDB Endowment, 2012, 5(9): 800 – 811.

[7] Judea P. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference[M]. San Fransisco: Morgan Kaufmann, 1989. 1 – 552.

[8] Jiawei H, Micheline K. Data Mining: Concepts and Techniques [M]. San Fransisco: Morgan Kaufmann, 2000. 204 – 206.

[9] Jie C, et al. Learning belief networks from data: an information theory based approach[A]. ACM International Conference on Information and Knowledge Management[C]. New York: ACM, 1997. 325 – 331.

作者简介

章登义　男,1965 年出生于湖北省荆州市.现为武汉大学计算机学院教授、博士生导师.
E-mail: dyzhangwhu@163.com

吴文李(通信作者)　女,1986 年 12 月出生于广东省湛江市.现为武汉大学博士研究生.主要研究方向为语义网数据挖掘.
E-mail: huihuigou@whu.edu.cn